DepressionMLP: A Multi-Layer Perceptron Architecture for Automatic Depression Level Prediction via Facial Keypoints and Action Units

Mingyue Niu, Ya Li, *Member, IEEE*, Jianhua Tao, *Senior Member, IEEE*, Xiuzhuang Zhou, *Member, IEEE*, and Björn W. Schuller, *Fellow, IEEE*

***Abstract*— Physiological studies have confirmed that there are differences in facial activities between depressed and healthy individuals. Therefore, while protecting the privacy of subjects, substantial efforts are made to predict the depression severity of individuals by analyzing Facial Keypoints Representation Sequences (FKRS) and Action Units Representation Sequences (AURS). However, those works has struggled to examine the spatial distribution and temporal changes of Facial Keypoints (FKs) and Action Units (AUs) simultaneously, which is limited in extracting the facial dynamics characterizing depressive cues. Besides, those works don't realize the complementarity of effective information extracted from FKRS and AURS, which reduces the prediction accuracy. To this end, we intend to use the recently proposed Multi-Layer Perceptrons with gating (gMLP) architecture to process FKRS and AURS for predicting depression levels. However, the channel projection in the gMLP disrupts the spatial distribution of FKs and AUs, leading to input and output sequences not having the same spatiotemporal attributes. This discrepancy hinders the additivity of residual connections in a physical sense. Therefore, we construct a novel MLP architecture named DepressionMLP. In this model, we propose the Dual Gating (DG) and Mutual Guidance (MG) modules. The DG module embeds cross-location and cross-frame gating results into the input sequence to maintain the physical properties of data to make up for the shortcomings of gMLP. The MG module takes the global information of FKRS (AURS) as a guidance mask to filter the AURS (FKRS) to achieve the interaction between FKRS and AURS. Experimental results on several benchmark datasets show the effectiveness of our method.**

***Index Terms*— Depression level prediction, facial keypoints, action units, dual gating module, mutual guidance module.**

This work was supported in part by the Basic Innovation and Scientific Research Cultivation Project of Yanshan University under Grant 2023LGQN006, in part by the National Natural Science Foun- dation of China under Grant 62271083 and Grant 61972046, in part by the Key Project of the National Language Commission under Grant ZDI145-81. This article was recommended by Associate Editor D. Gragnaniello. *(Corresponding author: Jianhua Tao.)*

Mingyue Niu is with the School of Information Science and Engineering, Yanshan University, Qinhuangdao, Hebei 066004, China, and also with the Key Laboratory for Computer Virtual Technology and System Integration of Hebei Province, Qinhuangdao 066004, China.

Ya Li and Xiuzhuang Zhou are with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: yli01@bupt.edu.cn; xiuzhuang.zhou@bupt.edu.cn).

Jianhua Tao is with the Department of Automation, Tsinghua University, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: jhtao@tsinghua.edu.cn).

Björn W. Schuller is with the ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, 86159 Augsburg, Germany, and also with the Group on Language, Audio, and Music (GLAM), Imperial College London, SW7 2BX London, U.K. (e-mail: bjoern.schuller@imperial.ac.uk).

Digital Object Identifier 10.1109/TCSVT.2024.3382334

I. Introduction

DEPRESSION can cause being troubled by negative emotions for a long time and affect one's physical and mental health. What is more serious is that patients with depression potentially tend to engage in self-mutilation and suicide due to the loss of interest in life [1]. Nowadays, due to the increase of uncertainties in society, people frequently have to bear more stress and thus have a higher chance of suffering from depression. The diagnosis of depression regularly requires the joint consultation of experienced doctors, which not only tires doctors but also delays the cure of patients. Facing the current situation, it is of practical value to use machine learning techniques to achieve automatic depression level prediction.

According to physiological studies [2], [3], the facial activities of depressed individuals differ from those of healthy individuals. That is to say, facial changes can be used as a biomarker to analyze the depression severity of an individual. Thus, many researchers [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15] have proposed various methods to capture depression cues from facial images or videos to automatically estimate the Beck Depression Inventory-II (BDI-II) scores [16], as shown in Table I. The works [4], [5] combine hand-crafted patterns with the Three Orthogonal Plane (TOP) framework [17] to obtain the histogram features, but those features are insensitive to salient information associated with depression [18]. Moreover, the process of extracting hand-crafted patterns is closely linked to the designer's experience, usually resulting in the loss of some valid information. To this end, some deep neural network models have been proposed to examine facial images [6], [7], [9], [10], [15] or videos [8], [11], [12], [13], [14] of individuals for extracting high-level representations of depression cues. However, collecting and processing the subjects' facial images and videos carries the hidden danger of privacy leakage.

TABLE I
THE BDI-II SCORE AND CORRESPONDING DEPRESSION DEGREES

BDI-II Score	Depression Degree
0–13	None
14–19	Mild
20–28	Moderate
29–63	Severe

Hence, some methods [19], [20], [21], [22], [23], [24], [25] utilize Facial Keypoint Representation Sequence (FKRS) and Action Units Representation Sequence (AURS) to predict the depression levels. However, both the Convolutional Neural Networks (CNN) used in [19], [20], [21], [22], and [24] and the Long Short Term Mermory (LSTM) used in [23] and [25] are difficult to simultaneously examine the spatial distribution and temporal changes of Facial Keypoints (FKs) and Action Units (AUs). Besides, those methods [19], [20], [21], [22], [23], [24], [25] don't explore the interaction between FKRS and AURS, which makes the effective information extracted from FKRS and AURS unable to complement each other and reduces the prediction accuracy.

Recently, the Multi-Layer Perceptrons (MLPs) with gating (gMLP) [26], a pure multi-layer perceptron architecture with less parameters, is designed and can be as good as Vision Transformer [27] in modeling sequences. However, when a sequence **A** is input into the channel projection in the gMLP, the resulting sequence **B** cannot have the same spatial structure as **A**. In this way, the residual connection of **A** and **B**, although numerically addable, are not guaranteed to be rational in a physical sense. Thus, the gMLP is not conducive to examining the spatial distribution differences of FKs and AUs among individuals with different depression levels.

To alleviate the above issues, we construct a novel MLP-based architecture termed as DepressionMLP to predict the BDI-II score using FKRS and AURS. In our model, we propose two novel modules namely Dual Gating (DG) and Mutual Guidance (MG) modules. The DG module embeds cross-location and cross-frame gating results into the input sequence through attention weighting to ensure that the input and output sequences have the same spatiotemporal attributes. Therefore, our DG module can compensate for the shortcomings of gMLP. Moreover, the MG module uses the global information of FKRS (AURS) as a guidance mask to filter AURS (FKRS) to achieve the interaction between two type of sequences and improve the perception of depression cues. It should be noted that although convolutional layers are used in our model, Conv1D and Conv2D with kernel size 1 are equivalent to right multiplication block matrices and matrix multiplication operations along channel dimensions, respectively. Therefore, our DepressionMLP still is an MLP architecture. In this way, we can follow the steps below to predict the depression severity of an individual:

Firstly, in the process of subjects completing interactive tasks, we use the OpenFace tooklit to record coordinates of FKs and AUs. The long-term FKRS and AURS can be obtained. Then, the long-term FKRS and AURS are divided into fixed-length short-term FKRS and AURS, which are fed into the constructed DepressionMLP architecture to predict the BDI-II scores. Finally, we take the average of BDI-II scores obtained from those short-term sequences as the prediction result corresponding to the subject. Experimental results on several publicly available depression databases demonstrate the effectiveness of our method.

Our main contribution can be summarized as follows:

(1) Our constructed DepressionMLP architecture enables end-to-end depression level prediction using FKRS and AURS. In addition, our DepressionMLP can obtain the prediction accuracy comparable to that of methods using facial images or videos.

(2) The proposed DG module compensates for the issues of gMLP's inability to maintain the spatiotemporal attributes of sequences and ensure the additivity of residual connections in the physical sense. Therefore, the DG module is more helpful in capturing facial dynamic patterns reflecting depression levels from FKRS and AURS.

(3) The proposed MG module can realize the interaction between two types of sequences in the form of mutual filtering. Therefore, the MG module is able to make the effective information extracted from FKRS and AURS complementary, and improve the model's perception of depression cues.

II. RELATED WORKS

In this paper, we propose a novel MLP architecture namely DepressionMLP to predict an individual's depression level. Therefore, in this section, we briefly review prior works on this subject.

A. Multi-Layer Perceptrons for Sequence Modeling

In recent years, some researchers [26], [28] questioned whether the convolution operation and attention mechanism are necessary conditions for visual tasks to achieve good experimental performance. Therefore, they used MLPs to build some simpler architectures to confirm their consideration. Tolstikhin et al. [28] proposed two types of MLP layers: one is the channel-mixing layer to independently process each token to achieve communication between channels, and the other is the token-mixing layer to independently process each channel to achieve communication between tokens. Furthermore, Liu et al. [26] proposed a network architecture of MLP with gating mechanism, which spat the input sequence into two independent parts along the channel dimension and realizes cross-token interaction through linear projection and gating function. However, Wang et al. [29] believed that the above MLP-based models all use static parameters to fuse tokens and are difficult to adapt to the content of the tokens to be fused. Hence, they presented a network architecture named DynaMixer, which generated a mixing matrix dependent on the input token sequence through two linear mappings and pumped it into the MLP-Mixer framework. Unlike the above works dealing with token sequences, Tang et al. [30] used MLPs along the row axis and column axis of the input tensor and obtained a sparse MLP model through parameter sharing.

From those above works, we can observe that MLP-based network architectures are effective for modeling sequence data. In other words, it is feasible to utilize MLPs to process FKRS and AURS to predict the depression level of an individual.

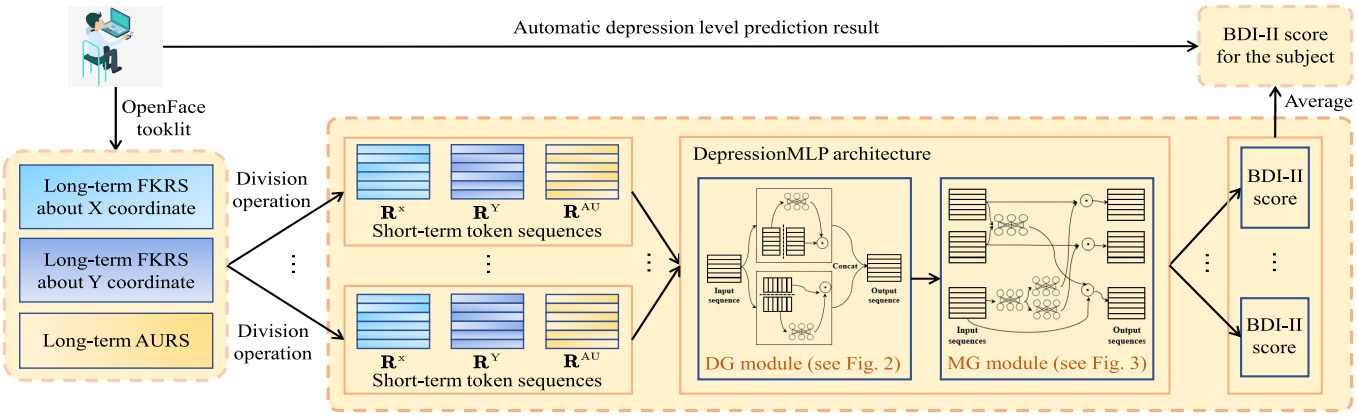


Fig. 1. The pipline of our DepressionMLP architecture for automatic depression level estimation. The OpenFace tooklit is used to extract the FKs and AUs. The FKRS (AURS) refers to the Facial Keypoints (Action Units) Representation Sequence. "$\mathbf{R}^{X}$" ("$\mathbf{R}^{Y}$") and "$\mathbf{R}^{AU}$" are obtained through dividing the long-term FKRS about X (Y) coordinate and the long-term AURS, respectively. "DG" and "MG" are short for Dual Gating and Mutual Guidance modules. "FCL" means to the Fully Connected Layer. The loss function is Root Mean Square Error (RMSE), as shown in Eq. (21).

B. Depression Level Prediction via Facial Features

In the past decade, many works [31], [32], [33] based on deep neural networks have been proposed and achieved good experimental performance. Thus, some researchers used network models to investigate the facial dynamics of individuals with different depression levels. Zhou et al. [7] constructed a DepressNet to select facial regions related to depression to predict BDI-II scores. Similarly, He et al. [34] presented a Deep Local Global Attention Convolutional Neural Network to highlight the facial regions related to depression. The authors in [12] proposed a Selective Element and Two Orders Vectorization (SE-TOV) network. In this network, the SE block was constructed to select useful elements from tensors generated with receptive fields of different sizes. Moreover, the TOV block was constructed to calculate first-order and second-order statistics of tensors for the purpose of tensor vectorization. Considering the spatiotemporal structure of facial changes, Melo et al. [35] constructed a Decomposed Multiscale Spatiotemporal Network (DMSN) to extract multi-scale representation of depression cues from spatial and temporal perspectives. Also to examine the spatiotemporal differences in the facial changes of individuals with different depression levels, Niu et al. [36] proposed a spatiotemporal attention mechanism to highlight discriminative frames related to depression for improving the prediction accuracy.

To protect the privacy of individuals, some methods based on FKs and AUs have been proposed. Syed et al. [22] used the coordinates of FKs to calculate the speed and acceleration of head movement, mouth opening, and eyelid movement. Next, they estimated an individual's depression level with the partial least squares method. Similarly, Yang et al. [19] calculated the range and speed of the displacements of FKs in the horizontal and vertical directions along the temporal axis to generate the histogram feature, which was fed into the CNN to predict an individual's depression level. Different from them, the authors in [20] adopted 3DCNN to encode FKRS and attention mechanism to weight the encoding results in spatial and temporal dimensions to capture facial dynamic patterns reflecting depression levels. Song et al. [24] used the Fourier Transform to convert AUs, head pose and gaze directions into spectral representations, which were input into CNN to estimate the depression severity of a subject. Furthermore, in the work of [21], Song et al. demonstrated how to use a fixed frequency set to obtain the spectral representation with the same size for any video. Again using multiple facial features, Ray et al. [23] used the LSTM to train regressors for predicting depression scores based on pose, gaze and AUs provided by the dataset, respectively. Moreover, Muzammel et al. [25] explored the impact of different fusion strategies of AUs and speech modalities on predicting individual depression levels.

As mentioned above, previous works make it difficult to examine both the spatial distribution and temporal changes of FKs or AUs of individuals with different depression levels. Besides, the depression-related information extracted from FKRS and AURS cannot be enhanced due to the lack of interaction between the two sequences.

III. Our Proposed DepressionMLP Architecture for Depression Level Prediction

In this paper, we construct a DepressionMLP architecture to predict the depression level using FKRS and AURS. In our method, the long-term FKRS and AURS are divided into short-term FKRS and AURS segments. Then, the DepressionMLP is used to estimate the BDI-II scores of those segments. Finally, the average of those BDI-II scores is taken as the assessment of an individual's depression level. Fig. 1 gives the complete process. As shown, "DG" and "MG" are the main two modules, which are presented in the following sections.

A. Short-Term Representation Sequences Generation

As mentioned before, we estimate an individual's depression severity using our constructed DepressionMLP model, which is an MLP architecture. Hence, it is necessary to generate the input token sequences. It is assumed that the OpenFace tooklit extracts FKs and AUs from all frames when the subject completes all interactive tasks. In this way, we can obtain the long-term FKRS about X coordinate, the long-term FKRS about Y coordinate and the long-term AURS. Next,

we divide these long-term representation sequences into fixed-length short-term representation sequences, which are denoted as "$\widehat{\mathbf{R}^{\mathrm{X}}}$", "$\widehat{\mathbf{R}^{\mathrm{Y}}}$" and "$\widehat{\mathbf{R}^{\mathrm{AU}}}$" as shown in Eq. (1).

$$\begin{cases} \widehat{\mathbf{R}^{\mathrm{X}}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1N} \\ \vdots & \vdots & & \vdots \\ x_{M1} & x_{M2} & \cdots & x_{MN} \end{bmatrix} \in \mathbb{R}^{M\times N} \\ \widehat{\mathbf{R}^{\mathrm{Y}}} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1N} \\ \vdots & \vdots & & \vdots \\ y_{M1} & y_{M2} & \cdots & y_{MN} \end{bmatrix} \in \mathbb{R}^{M\times N} \\ \widehat{\mathbf{R}^{\mathrm{AU}}} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1P} \\ \vdots & \vdots & & \vdots \\ a_{M1} & a_{M2} & \cdots & a_{MP} \end{bmatrix} \in \mathbb{R}^{M\times P} \end{cases}, \tag{1}$$

where M is the number of frames for short-term representation sequences. $N = 68$ and $P = 35$ are the number of FKs and AUs, respectively.

To make the values in $\widehat{\mathbf{R}^{\mathrm{X}}}$ and $\widehat{\mathbf{R}^{\mathrm{Y}}}$ within the range of [0, 1], we use Eq. (2) to normalize them. The corresponding results are denoted as $\mathbf{R}^{\mathrm{X}} \in \mathbb{R}^{M\times N}$ and $\mathbf{R}^{\mathrm{Y}} \in \mathbb{R}^{M\times N}$. Due to the splitting operation in the DG module, the number of rows and columns in $\widehat{\mathbf{R}^{\mathrm{AU}}}$ must be even. So, we use Eq. (3) to interpolate $\widehat{\mathbf{R}^{\mathrm{AU}}}$ to obtain $\mathbf{R}^{\mathrm{AU}} \in \mathbb{R}^{M\times N}$.

$$\begin{cases} \mathbf{R}^{\mathrm{X}} = \widehat{\mathbf{R}^{\mathrm{X}}} / \left(\max\left(I_r, I_c\right)\right) \in \mathbb{R}^{M\times N} \\ \mathbf{R}^{\mathrm{Y}} = \widehat{\mathbf{R}^{\mathrm{Y}}} / \left(\max\left(I_r, I_c\right)\right) \in \mathbb{R}^{M\times N} \end{cases}, \tag{2}$$

where I_r, I_c are the number of rows and columns of the frame, respectively. $\max(\cdot,\cdot)$ refers to taking the maximum value of both. "/" refers to dividing a matrix by a constant.

$$\mathbf{R}^{\mathrm{AU}}\left[:, j\right] = \begin{cases} \left(\widehat{\mathbf{R}^{\mathrm{AU}}}\left[:, \dfrac{j+1}{2}\right]\right), & j \text{ is an odd} \\ \dfrac{\left(\widehat{\mathbf{R}^{\mathrm{AU}}}\left[:, \frac{j}{2}\right] + \widehat{\mathbf{R}^{\mathrm{AU}}}\left[:, \frac{j}{2}+1\right]\right)}{2}, & j \text{ is an even,} \end{cases} \tag{3}$$

where $\mathbf{R}^{\mathrm{AU}}\left[:, j\right]$ refers to the j-th column of $\mathbf{R}^{\mathrm{AU}}$ and $j = 1, \cdots, N$.

From the above process, it is easy to observe that "$\mathbf{R}^{\mathrm{X}}$", "$\mathbf{R}^{\mathrm{Y}}$" and "$\mathbf{R}^{\mathrm{AU}}$" are essentially temporal. Meanwhile, those sequences also reflect facial changes, which is beneficial for predicting depression levels.

B. Dual Gating Module for Capturing Depression Cues

To extract depression cues from FKRS and AURS, we propose a DG module to compensate for the limitation of gMLP. Thus, in this section, we present the architecture of gMLP and analyze its shortcomings in modeling temporal sequences. Then, a detailed description of DG module is provided.

1) MLPs Layer With Gating: As a pure MLP architecture, the gMLP is composed of several identical blocks. For simplicity, we still refer to each block as gMLP. The purpose of gMLP is to use the cross-token interaction to gate the cross-channel interaction. Actually, the gMLP uses Eq. (4) to achieve channel projection of the input sequence $\mathbf{X} \in \mathbb{R}^{L\times C}$, where L and C denote the length of the sequence and the number of channels, respectively. The result is denoted as $\mathbf{Z}$. Then, along the channel dimension, $\mathbf{Z}$ is spit into two parts $(\mathbf{Z}_1, \mathbf{Z}_2)$ with the same number of channels. Eq. (5) is adopted to obtain the gating result $\hat{\mathbf{Z}}$. Finally, the output $\mathbf{G}$ of gMLP is gained via channel projection and residual connection as in Eq. (6).

$$\mathbf{Z} = \gamma\left(\mathcal{N}_{\mathrm{C}}(\mathbf{X}) \otimes \mathbf{W}_1\right) \in \mathbb{R}^{L\times D}, \tag{4}$$

where $\mathbf{W}_1 \in \mathbb{R}^{C\times D}$ is a learnable parameter matrix. $\mathcal{N}_{\mathrm{C}}(\cdot)$ refers to the normalization operation along the channel dimension of the sequence. $\gamma(\cdot)$ is the nonlinear activation function GELU [37]. "$\otimes$" represents the matrix multiplication operation.

$$\hat{\mathbf{Z}} = \mathbf{Z}_1 \odot \left(\mathbf{W}_2 \otimes \mathbf{Z}_2\right) \in \mathbb{R}^{L\times (D/2)}, \tag{5}$$

where $\mathbf{W}_2 \in \mathbb{R}^{L\times L}$ is a learnable parameter matrix. "$\odot$" refers to element-wise multiplication.

$$\mathbf{G} = \mathbf{X} \oplus \left(\hat{\mathbf{Z}} \otimes \mathbf{W}_3\right) \in \mathbb{R}^{L\times C}, \tag{6}$$

where $\mathbf{W}_3 \in \mathbb{R}^{(D/2)\times C}$ is a learnable parameter matrix. "$\oplus$" is an element-wise addition operation.

From the above description, it can be seen that the gMLP aggregates different channels through matrix right multiplication during the channel projection process as in Eq. (4). However, for the $\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$ and $\mathbf{R}^{\mathrm{AU}}$, this operation makes the number of columns in the processed sequence not necessarily equal to the number of FKs or AUs. In other words, Eq. (4) cannot make the processed sequence have the same spatial structure as $\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$ and $\mathbf{R}^{\mathrm{AU}}$. Therefore, the gMLP is limited in capturing the differences in facial changes among individuals with different depression levels.

2) The Dual Gating Module: To alleviate those issues in the gMLP, we propose the DG module to embed cross-location and cross-frame gating results into the $\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$ and $\mathbf{R}^{\mathrm{AU}}$ in the form of weights to ensure that the output sequences can maintain the same spatiotemporal attributes as the input sequences. Fig. 2 displays the flow of DG module. As shown, the DG module contains two parts: cross-location gating embedding and cross-frame gating embedding. Since $\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$ and $\mathbf{R}^{\mathrm{AU}}$ have the same size, we use $\mathbf{R} \in \mathbb{R}^{M\times N}$ instead of $\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$ and $\mathbf{R}^{\mathrm{AU}}$ for brevity of description.

For the cross-location gating embedding, we use Eq. (7) to summarize the spatial structure of $\mathbf{R}$. The corresponding result is recorded as $\mathbf{R}^{\mathrm{L}}$. It is necessary to point out that the Conv1D layer in Fig. 2 and Eq. (7) is equivalent to right multiplying a parameter matrix for $\mathbf{R}$. In this way, each column of $\mathbf{R}^{\mathrm{L}}$ can be regarded as the aggregation result of all columns of $\mathbf{R}$. In other words, each column of $\mathbf{R}^{\mathrm{L}}$ has the same physical meaning. Thus, similar to the gMLP, $\mathbf{R}^{\mathrm{L}}$ can be split into two parts $\mathbf{R}^{\mathrm{L}}_1 \in \mathbb{R}^{M\times \frac{N}{2}}$ and $\mathbf{R}^{\mathrm{L}}_2 \in \mathbb{R}^{M\times \frac{N}{2}}$ of the same size along the column axis. Immediately afterwards, the Eq. (8) is used to capture information about the cross-frame interaction. The result is recorded as $\mathbf{M}^{\mathrm{F}}_2$. And we use Eq. (9) to implement gating operation. The corresponding result is denoted as $\mathbf{G}^{\mathrm{F}\to\mathrm{L}}$. To preserve the spatiotemporal attribute of the sequence, we use the attention mechanism to embed $\mathbf{G}^{\mathrm{F}\to\mathrm{L}}$ into $\mathbf{R}$ and obtain $\mathbf{R}^{\mathrm{F}\to\mathrm{L}}$ as in Eq. (10).

$$\mathbf{R}^{\mathrm{L}} = \delta\left(\mathcal{C}_1(\mathbf{R})\right) \in \mathbb{R}^{M\times N}, \tag{7}$$

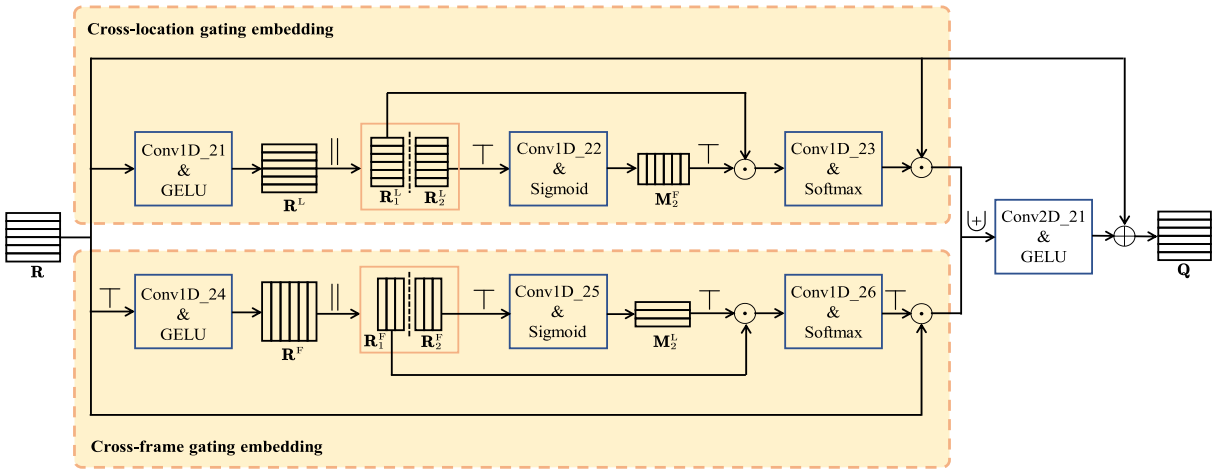


Fig. 2. The flow of DG module. "$\|$", "$\top$" and "$\uplus$" are split, matrix transpose and concatenation operations, respectively. "$\odot$" and "$\oplus$" refer to element-wise multiplication and element-wise addition, respectively. For simplicity, "$\mathbf{R}$" represents "$\mathbf{R}^{\mathrm{X}}$", "$\mathbf{R}^{\mathrm{Y}}$" or "$\mathbf{R}^{\mathrm{AU}}$". "$\mathbf{Q}$" is the output result of "$\mathbf{R}$" processed by the DG module. Thus, "$\mathbf{Q}$" can be "$\mathbf{Q}^{\mathrm{X}}$", "$\mathbf{Q}^{\mathrm{Y}}$" or "$\mathbf{Q}^{\mathrm{AU}}$".

where $\mathcal{C}_1(\cdot)$ is a Conv1D layer and $\delta(\cdot)$ is the ReLU activation function.

$$\mathbf{M}_2^{\mathrm{F}} = \sigma(\mathcal{C}_1((\mathbf{R}_2^{\mathrm{L}})^\top)) \in \mathbb{R}^{\frac{N}{2}\times M}, \tag{8}$$

where $\mathbf{R}_2^{\mathrm{L}} \in \mathbb{R}^{M\times\frac{N}{2}}$ is obtained by splitting $\mathbf{R}^{\mathrm{L}}$ into two parts of the same size i.e., $\mathbf{R}_1^{\mathrm{L}} \in \mathbb{R}^{M\times\frac{N}{2}}$ and $\mathbf{R}_2^{\mathrm{L}} \in \mathbb{R}^{M\times\frac{N}{2}}$. $\sigma(\cdot)$ is the Sigmoid activation function.

$$\mathbf{G}^{\mathrm{F}\to\mathrm{L}} = (\mathbf{M}_2^{\mathrm{F}})^\top \odot \mathbf{R}_1^{\mathrm{L}} \in \mathbb{R}^{M\times\frac{N}{2}}. \tag{9}$$

$$\mathbf{R}^{\mathrm{F}\to\mathrm{L}} = \mathtt{Softmax}(\mathcal{C}_1(\mathbf{G}^{\mathrm{F}\to\mathrm{L}})) \odot \mathbf{R} \in \mathbb{R}^{M\times N}. \tag{10}$$

For the cross-frame gating embedding, we use Eq. (11) instead of Eq. (7) to summarize the temporal changes of $\mathbf{R}$. The corresponding result is denoted as $\mathbf{R}^{\mathrm{F}}$. Therefore, in the same way as $\mathbf{R}^{\mathrm{L}}$, each column of $\mathbf{R}^{\mathrm{F}}$ has the same physical meaning. In this way, we split $\mathbf{R}^{\mathrm{F}}$ into the same two parts $\mathbf{R}_1^{\mathrm{F}} \in \mathbb{R}^{N\times\frac{M}{2}}$ and $\mathbf{R}_2^{\mathrm{F}} \in \mathbb{R}^{N\times\frac{M}{2}}$. Moreover, we replace the $\mathbf{R}_2^{\mathrm{L}}$ in Eq. (8) with $\mathbf{R}_2^{\mathrm{F}}$ to obtain the cross-location interaction denoted as $\mathbf{M}_2^{\mathrm{L}}$. Therefore, we can implement the gating operation as in Eq. (12). Similarly, in order to maintain the spatiotemporal attributes of the sequence, we use $(\mathbf{G}^{\mathrm{L}\to\mathrm{F}})^\top$ instead of $\mathbf{G}^{\mathrm{F}\to\mathrm{L}}$ in Eq. (10) to embed $\mathbf{G}^{\mathrm{L}\to\mathrm{F}}$ into $\mathbf{R}$ and denote the result as $\mathbf{R}^{\mathrm{L}\to\mathrm{F}}$. Finally, Eq. (13) is used to combine $\mathbf{R}^{\mathrm{F}\to\mathrm{L}}$ and $\mathbf{R}^{\mathrm{L}\to\mathrm{F}}$ and obtain the output $\mathbf{Q} \in \mathbb{R}^{M\times N}$ of DG module by residual connection.

$$\mathbf{R}^{\mathrm{F}} = \delta(\mathcal{C}_1(\mathbf{R}^\top)) \in \mathbb{R}^{N\times M}. \tag{11}$$

$$\mathbf{G}^{\mathrm{L}\to\mathrm{F}} = (\mathbf{M}_2^{\mathrm{L}})^\top \odot \mathbf{R}_1^{\mathrm{F}} \in \mathbb{R}^{N\times\frac{M}{2}}. \tag{12}$$

$$\mathbf{Q} = \delta(\mathcal{C}_2(\mathtt{Concat}(\mathbf{R}^{\mathrm{F}\to\mathrm{L}}, \mathbf{R}^{\mathrm{L}\to\mathrm{F}}))) \oplus \mathbf{R} \in \mathbb{R}^{M\times N}, \tag{13}$$

where $\mathcal{C}_2(\cdot)$ is Conv2D layer and $\mathtt{Concat}(\cdot,\cdot)$ refers to the concatenation operation along the channel axis.

From the above description, we can easily see that on the one hand, the DG module utilizes the interaction among frames (location) to implement the gating operation on cross-location (cross-frame) aggregation results to select information related to depression. On the other hand, the DG module adopts the attention mechanism to ensure that the input and output sequences have the same spatiotemporal attributes. In this way, our proposed DG module not only enhances the discriminative ability of depression cue representation, but also guarantees the additivity of residual connections, thus compensating for the limitations of gMLP.

C. The Mutual Guidance Module for Sequence Interaction

As mentioned above, we select the information associated with depression in $\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$ and $\mathbf{R}^{\mathrm{AU}}$ using the DG module and obtain the $\mathbf{Q}^{\mathrm{X}}$, $\mathbf{Q}^{\mathrm{Y}}$ and $\mathbf{Q}^{\mathrm{AU}}$. However, FKRS and AURS are two types of sequences to characterize facial dynamics. Therefore, it is necessary to take advantages of $(\mathbf{Q}^{\mathrm{X}}, \mathbf{Q}^{\mathrm{Y}})$ and $\mathbf{Q}^{\mathrm{AU}}$ to enhance the model's representation ability for depression cues. To this end, we propose the MG module to take the global information of $(\mathbf{Q}^{\mathrm{X}}, \mathbf{Q}^{\mathrm{Y}})$ or $\mathbf{Q}^{\mathrm{AU}}$ as a guidance mask to filter $\mathbf{Q}^{\mathrm{AU}}$ or $(\mathbf{Q}^{\mathrm{X}}, \mathbf{Q}^{\mathrm{Y}})$ to achieve the interaction of effective information extracted from FKRS and AURS. Fig. 3 gives the flow of MG module.

For the FKRS's guidance to AURS, the Eq. (14) summarizes the temporal changes for each facial keypoint in $\mathbf{Q}^{\mathrm{X}}$ and $\mathbf{Q}^{\mathrm{Y}}$. The results are denoted as $\mathbf{q}^{\mathrm{X}}$ and $\mathbf{q}^{\mathrm{Y}}$. Moreover, as shown in Eq. (15), we integrate $\mathbf{q}^{\mathrm{X}}$ and $\mathbf{q}^{\mathrm{Y}}$, the result is denoted as $\mathbf{q}^{\mathrm{XY}}$. After that, Eq. (16) is used to implement the FKRS's guidance to each facial action unit. Note that the term of $\sigma(\mathcal{C}_1(\mathbf{q}^{\mathrm{XY}})) \in \mathbb{R}^{M\times N}$ in Eq. (16) is the guidance mask in the guidance process of the FKRS to AURS.

$$\begin{cases} \mathbf{q}^{\mathrm{X}} = \delta(\mathcal{C}_1((\mathbf{Q}^{\mathrm{X}})^\top)) \in \mathbb{R}^{N\times 1} \\ \mathbf{q}^{\mathrm{Y}} = \delta(\mathcal{C}_1((\mathbf{Q}^{\mathrm{Y}})^\top)) \in \mathbb{R}^{N\times 1}. \end{cases} \tag{14}$$

$$\mathbf{q}^{\mathrm{XY}} = \delta(\mathbf{W}^{\mathrm{XY}} \otimes \mathtt{Concat}(\mathbf{q}^{\mathrm{X}}, \mathbf{q}^{\mathrm{Y}})) \in \mathbb{R}^{M\times 1}, \tag{15}$$

where $\mathbf{W}^{\mathrm{XY}} \in \mathbb{R}^{M\times 2N}$ is a learnable parameter matrix i.e., a fully connected layer. $\mathtt{Concat}(\cdot,\cdot)$ refers to the concatenation operation along the row axis.

$$\mathbf{D}^{\mathrm{P}\to\mathrm{U}} = \sigma(\mathcal{C}_1(\mathbf{q}^{\mathrm{XY}})) \odot \mathbf{Q}^{\mathrm{AU}} \in \mathbb{R}^{M\times N}. \tag{16}$$

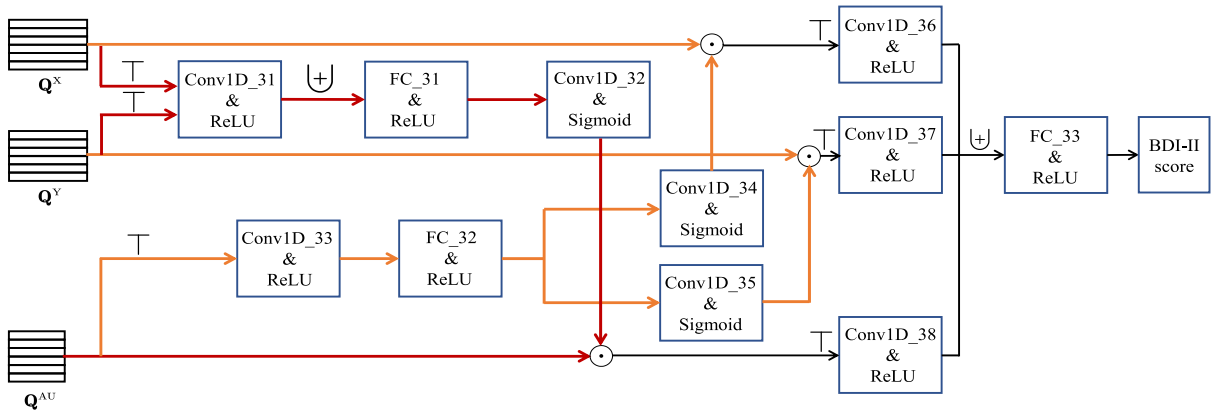


Fig. 3. The flow of MG module. "$\mathbf{Q}^{\mathrm{X}}$", "$\mathbf{Q}^{\mathrm{Y}}$" and "$\mathbf{Q}^{\mathrm{AU}}$" are obtained by inputting "$\mathbf{R}^{\mathrm{X}}$", "$\mathbf{R}^{\mathrm{Y}}$" and "$\mathbf{R}^{\mathrm{AU}}$" to the DM module. "$\top$" and "$\uplus$" are matrix transpose and concatenation operations, respectively. "$\odot$" refers to element-wise multiplication. The red (orange) arrow indicates the guidance process of the FKRS (AURS) to AURS (FKRS).

For the AURS's guidance to FKRS, we use Eq. (17) to summarize each facial action unit and obtain $\mathbf{q}^{\mathrm{AU}}$. Immediately after, Eq. (18) is adopted to implement the AURS's guidance to FKRS. The corresponding results are referred to as $\mathbf{D}^{\mathrm{U}\to\mathrm{X}}$ and $\mathbf{D}^{\mathrm{U}\to\mathrm{Y}}$. Note that the term of $\sigma(\mathcal{C}_1(\mathbf{q}^{\mathrm{AU}})) \in \mathbb{R}^{M\times N}$ in Eq. (18) is the guidance mask in the guidance process of AURS to FKRS.

$$\mathbf{q}^{\mathrm{AU}} = \delta(\mathbf{W}^{\mathrm{AU}} \otimes \delta(\mathcal{C}_1((\mathbf{Q}^{\mathrm{AU}})^\top))) \in \mathbb{R}^{M\times 1}, \tag{17}$$

where $\mathbf{W}^{\mathrm{AU}} \in \mathbb{R}^{M\times N}$ is a learnable parameter matrix i.e., a fully connected layer.

$$\begin{cases} \mathbf{D}^{\mathrm{U}\to\mathrm{X}} = \sigma(\mathcal{C}_1(\mathbf{q}^{\mathrm{AU}})) \odot \mathbf{Q}^{\mathrm{X}} \in \mathbb{R}^{M\times N} \\ \mathbf{D}^{\mathrm{U}\to\mathrm{Y}} = \sigma(\mathcal{C}_1(\mathbf{q}^{\mathrm{AU}})) \odot \mathbf{Q}^{\mathrm{Y}} \in \mathbb{R}^{M\times N}. \end{cases} \tag{18}$$

After completing the mutual guidance between FKRS and AURS, we use Eq. (19) to calculate the estimated BDI-II score for each facial keypoint and action unit. The corresponding results are regarded as $\mathbf{s}^{\mathrm{X}}$, $\mathbf{s}^{\mathrm{Y}}$ and $\mathbf{s}^{\mathrm{AU}}$. Moreover, Eq. (20) is used to integrate these results to obtain the predicted value for the BDI-II score.

$$\begin{cases} \mathbf{s}^{\mathrm{X}} = \delta(\mathcal{C}_1((\mathbf{D}^{\mathrm{U}\to\mathrm{X}})^\top)) \in \mathbb{R}^{N\times 1} \\ \mathbf{s}^{\mathrm{Y}} = \delta(\mathcal{C}_1((\mathbf{D}^{\mathrm{U}\to\mathrm{Y}})^\top)) \in \mathbb{R}^{N\times 1} \\ \mathbf{s}^{\mathrm{AU}} = \delta(\mathcal{C}_1((\mathbf{D}^{\mathrm{P}\to\mathrm{U}})^\top)) \in \mathbb{R}^{N\times 1}. \end{cases} \tag{19}$$

$$s = \delta(\mathbf{W} \otimes \mathtt{Concat}(\mathbf{s}^{\mathrm{X}}, \mathbf{s}^{\mathrm{Y}}, \mathbf{s}^{\mathrm{AU}})) \in \mathbb{R}, \tag{20}$$

where $\mathbf{W} \in \mathbb{R}^{1\times 3N}$ is a learnable parameter matrix i.e., a fully connected layer. $\mathtt{Concat}(\cdot,\cdot)$ refers to the concatenation operation along the row axis.

From the above process, on the one hand, the MG module achieves the interaction among two types of facial representation sequences through Eqs. (16) and (18). On the other hand, it utilizes ($\mathbf{Q}^{\mathrm{X}}$ and $\mathbf{Q}^{\mathrm{Y}}$) or $\mathbf{Q}^{\mathrm{AU}}$ to generate the guidance mask to filter irrelevant parts in $\mathbf{Q}^{\mathrm{AU}}$ or ($\mathbf{Q}^{\mathrm{X}}$ and $\mathbf{Q}^{\mathrm{Y}}$) while retaining the discriminative information, which can reflect the facial dynamic differences of individuals with different depression levels. In addition, during the BDI-II score estimation stage, we calculate and integrate the result of each facial keypoint and action unit, which can comprehensively examine the role of each facial component in the depression level estimation task.

The relationship between Cross-Attention and MG module: The Cross-Attention (CA) mechanism is a widely used method for fusing two sequences (set as $\mathbf{S}_1$ and $\mathbf{S}_2$). The core of CA is to aggregate $\mathbf{S}_1$ using the similarity results of $\mathbf{S}_1$ and $\mathbf{S}_2$. Thus, the result of CA no longer have the same temporal attribute as $\mathbf{S}_1$. Unlike this, our proposed MG module uses the global information of $\mathbf{S}_1$ to generate a guidance mask for filtering $\mathbf{S}_2$ element by element. Therefore, the output of MG module has the same temporal attribute as the input sequence, which is more conductive to capturing discriminative facial dynamics.

IV. Experiments

In this section, we briefly describe several databases used in our experiments and present the details of our model. Then, some ablation experiments are performed to show the feasibility of proposed modules. Finally, we illustrate the effectiveness of our method by comparing it with previous works.

A. Databases and Metrics

In 2013 and 2014, Valstar et al. initiated a sub-track using audio and video modalities to predict individual depression levels in the Audio/Video Emotion Challenge (AVEC) competition and released the corresponding databases.

The AVEC 2013 depression database collects audio and video modal signals generated by participants during the completion of human-computer interaction tasks. Specifically, participants are required to sit in front of a computer and perform 14 different tasks based on on-screen prompts. All participants are native German speakers and between the ages of 18 and 63. In the AVEC 2013 databases, there are 150 video recordings from 82 participants. The duration of these 150 samples ranges from 20 minutes to 50 minutes. And these 150 recordings are equally divided by the publisher into training, development and test sets. It should be noted that each participant is required to fill out the BDI-II scale before conducting human-computer interaction tasks, and the scale score is considered as the label for the corresponding participant.

The AVEC 2014 depression database is taken over from the AVEC 2013 competition. Thus, they have the same acquisition settings and participant age distribution. However, there are only two tasks included in the AVEC 2014 database. In each task, 150 videos are collected with durations ranging from 6 seconds to 4 minutes 8 seconds. And these 150 videos are equally divided by the publisher into training, development and test sets. Note that, in our experiments, we combine the training, development, and test sets of those two tasks, respectively. Hence, there are 100 video recordings in the training, development, and test sets. While, we still refer to the merged database as AVEC 2014.

To measure the accuracy of depression level estimation, researchers often adopt the RMSE and MAE metrics. The RMSE and MAE are calculated as shown in Eq. (21) and Eq. (22), respectively. In addition, we also employ R^2 and Symmetric Mean Absolute Percentage Error (SMAPE) metrics in the ablation experiments. The calculation formulas for them are given in Eqs. (23) and (24).

$$\mathrm{RMSE} = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(y_k - \widehat{y_k})^2}, \tag{21}$$

where K is the number of samples. y_k and $\widehat{y_k}$ are the true and estimated BDI-II scores of the k-th sample, respectively.

$$\mathrm{MAE} = \frac{1}{K}\sum_{k=1}^{K}|y_k - \widehat{y_k}|. \tag{22}$$

$$R^2 = 1 - \frac{\sum_{k=1}^{K}(y_k - \widehat{y_k})^2}{\sum_{k=1}^{K}(y_k - \overline{y})^2}, \tag{23}$$

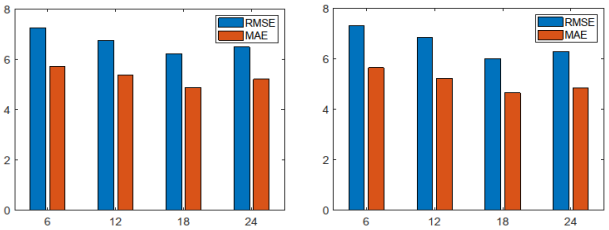
where $\overline{y}$ is the mean of true values of the K samples.

$$\mathrm{SMAPE} = \frac{100\%}{K}\sum_{k=1}^{K}\frac{|y_k - \widehat{y_k}|}{(|y_k| + |\widehat{y_k}|)/2}. \tag{24}$$

In addition, to demonstrate the robustness of our method, we also conduct experiments on the Distress Analysis Interview Corpus/Wizard of Oz (DAIC-WOZ) dataset [38]. Unlike the AVEC 2013 and AVEC 2014 datasets, the DAIC-WOZ does not provide the original videos and only provides 68 FKs and 20 AUs for privacy reasons. The DAIC-WOZ dataset is the part of the Distress Analysis Interview Corpus(DAIC), which supports the diagnosis of psychological states such as anxiety, depression and post-traumatic stress disorder through designed clinical interviews. During the interview, the facial movements of the subjects are recorded. For the DAIC-WOZ dataset, it contains 189 participants, which answer the questions of an animated virtual interviewer name Ellie. The label of each participant is the PHQ-8 score. For the PHQ-8 score, 0-4 is none, 5-9 is mild, 10-14 is moderate, 15-19 is moderate to severe and 20-24 is severe. It should be noted that the experiments on DAIC-WOZ is to demonstrate the robustness of our method to different depression scale scores. Therefore, we only present the comparative results in Part *D*.

B. Experimental Settings

1) Data Preprocessing: Some researchers [39], [40] have found that the best prediction accuracy is gained when the frame rate is reduced from 30fps to 6fps. Hence, the frame rate is adjusted to 6fps for the AVEC 2013 and AVEC 2014 databases, and the resolution of all frames in a video segment has been rescaled to 480 × 480. Moreover, we employ the OpenFace toolkit [43] to extract the 68 FKs and 35 AUs of the face. To determine the value of M in Eq. (1), we perform experiments on the development sets of AVEC 2013 and AVEC 2014. The corresponding results are shown in Fig. 4. Therefore, we take the value of M to be 18, which corresponds to 18 consecutive video frames in a short-term video segment. In this way, according to Eqs. (1)-(3), a short-term video segment with 18 frames can be encoded as three token sequences, namely $\mathbf{R}^{\mathrm{X}} \in \mathbb{R}^{18\times 68}$, $\mathbf{R}^{\mathrm{Y}} \in \mathbb{R}^{18\times 68}$ and $\mathbf{R}^{\mathrm{AU}} \in \mathbb{R}^{18\times 68}$. It should be pointed out that the label of a short-term video segment is the same as the label of the corresponding long-term video. And the coverage rate between two adjacent video segments is 50%.



(a) Prediction accuracy on AVEC 2013 development set for different M values.

(b) Prediction accuracy on AVEC 2014 development set for different M values.

Fig. 4. The impact of different M values in Eq. (1) on the prediction accuracy of depression level on the AVEC 2013 (a) and AVEC 2014 (b) development sets.

TABLE II
NETWORK LAYER PARAMETER SETTINGS FOR OUR PROPOSED MODULES

Modules	Layers	#Filters (#Neurons)	Kernel Size
DG Module (Fig. 2)	Conv1D_21	68	1
	Conv1D_22	18	1
	Conv1D_23	68	1
	Conv1D_24	18	1
	Conv1D_25	68	1
	Conv1D_26	18	1
	Conv2D_21	1	1 × 1
MG Module (Fig. 3)	Conv1D_31	1	1
	FC_31	18	/
	Conv1D_32	68	3
	Conv1D_33	1	1
	FC_32	18	/
	Conv1D_34	68	3
	Conv1D_35	68	3
	Conv1D_36	1	1
	Conv1D_37	1	1
	Conv1D_38	1	1
	FC_33	1	/

2) The Architecture of the DepressionMLP Model: In this paper, as shown in Fig. 1, the DepressionMLP architecture is constructed to estimate the depression level of an individual with FKs and AUs. In this model, we propose DG and MG modules, which are shown in Fig. 2 and Fig. 3, respectively. The corresponding network layer parameter settings are presented in Table II.

3) Model Training: As pointed in [41], the data scheduling is important for model training. In the AVEC 2013 and AVEC 2014 datasets, the distribution of subjects' labels

TABLE III

EXPERIMENTAL PERFORMANCE USING DIFFERENT MODELS AND INPUTS ON AVEC 2013 AND AVEC 2014 DEVELOPMENT SETS. "CLG" AND "CFG" REFER TO CROSS-LOCATION GATING EMBEDDING AND CROSS-FRAME GATING EMBEDDING, RESPECTIVELY

Models	Input(s)	AVEC 2013				AVEC 2014			
		RMSE (↓)	MAE (↓)	R^2 (↑)	SMAPE (↓)	RMSE (↓)	MAE (↓)	R^2 (↑)	SMAPE (↓)
gMLP+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$	9.10	7.45	0.42	63.21%	9.08	7.32	0.42	72.01%
	$\mathbf{R}^{\mathrm{AU}}$	8.78	7.04	0.46	60.97%	8.43	6.79	0.50	67.85%
	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	8.49	6.87	0.49	61.49%	8.18	6.63	0.53	66.73%
CLG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$	8.87	7.14	0.45	64.72%	8.80	6.83	0.46	61.04%
	$\mathbf{R}^{\mathrm{AU}}$	8.73	6.61	0.46	57.53%	8.52	6.48	0.49	60.84%
	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	8.29	6.33	0.52	64.37%	8.07	6.28	0.54	61.70%
CFG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$	9.05	7.30	0.42	72.00%	8.91	7.06	0.44	67.73%
	$\mathbf{R}^{\mathrm{AU}}$	8.81	7.46	0.45	67.30%	8.71	6.85	0.47	61.69%
	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	8.48	6.84	0.49	69.44%	8.40	6.72	0.50	59.78%
DG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$	8.53	6.81	0.49	60.29%	8.74	7.07	0.46	63.48%
	$\mathbf{R}^{\mathrm{AU}}$	8.38	6.65	0.50	63.76%	8.61	6.96	0.48	65.09%
	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	**8.16**	**6.28**	**0.53**	**56.01%**	**7.99**	**6.16**	**0.55**	**62.44%**

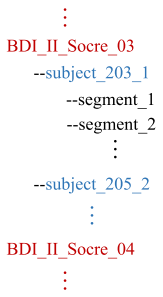


Fig. 5. Data storage form for AVEC 2013 and AVEC 2014 training sets.

(i.e. BDI-II scores) is uneven. Meanwhile, to eliminate the influence of subject identity on predicting depression levels, we store the training data in the manner shown in Fig. 5. In this way, we take 4 segments from the file of each subject (e.g., subject_203_1) under each score (e.g., BDI_II_Socre_03) to form an epoch (the next epoch is similar) and set the batchsize to 8 to train our model. For our model, the loss function is RMSE as shown in Eq. (21). We use the Keras deep learning framework and adopt the Adam [42] optimizer with default momentum values (0.9,0.999) for (β_1 and β_2). And the weight decay and initial learning rate are set to 0.0001 and 0.0002, respectively.

4) Model Validation and Testing: We use development sets and test sets to complete the validation and testing of the model. In both phases, the development and test sets are preprocessed as described in Section IV *B 1)*. Then, the average of the predicted scores of all video segments is used as the estimate of the corresponding subject's BDI-II score.

C. Analysis of Ablation Experiments

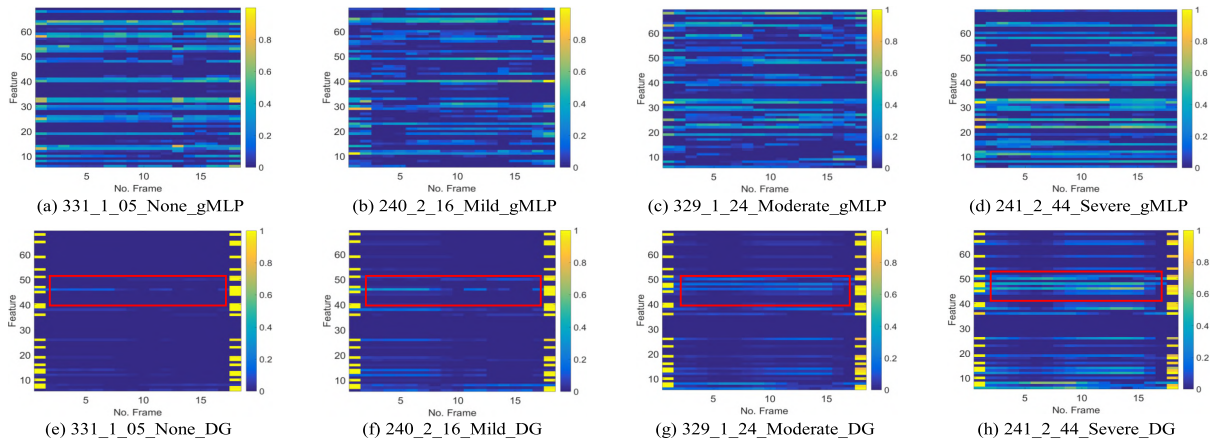
As described in Section III, our constructed DepressionMLP model mainly contains the DG and MG modules. Hence, in this section, we perform some ablation experiments on the development databases of AVEC 2013 and AVEC 2014 to clarify the effectiveness of these two modules.

1) The Role of DG Module: We propose the DG module to capture depression cues contained in FKRS and AURS through cross-location and cross-frame gating embedding processes. Thus, some experiments are conducted to demonstrate the effectiveness of DG module in estimating BDI-II scores. The corresponding results are shown in Table III.

From Table III, we can observe that the experimental performance of "CLG+Linear" is better that of "gMLP+Linear". This is because the gMLP module cannot ensure the temporal attribute of input sequences and the additivity of residual connection. In contrast, the cross-location gating process adopts the attention mechanism to embed the gating result into the input sequence, which not only guarantees the rationality of data attributes and calculations, but also highlights the information related to depression. This reason can also be used to explain the comparison of "gMLP+Linear" and "CFG+Linear".

Moreover, it is not difficult to find from the comparison between "CLG+Linear" and "CFG+Linear" that the cross-frame gating process is more helpful in capturing facial dynamic differences among individuals with different depression levels than the cross-location gating process. The reason is that the cross-frame gating process examines the temporal changes of each FK, which can reflects the impact of depression disorder on individual facial activity. The cross-location gating process focuses on examining the spatial distribution of FKs, so it is limited in extracting facial dynamic features. Moreover, this fact indicates that it is beneficial to jointly explore the temporal changes and spatial distribution of FKs for capturing depression cues.

In addition, from the Table III, we can find that inputting $\mathbf{R}^{\mathrm{AU}}$ can obtain better prediction accuracy than ($\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$). This result can be explained that AUs can characterize changes in facial muscle movement, which helps reveal the impact of depression disorder on individual facial activities. FKs describe the geometric structure of the face, but some difficulties arise in extracting depression cues due to the spatial distribution of individual facial organs [43]. Moreover, one can find that joint examination of FKs and AUs can improve the experimental performance of depression level prediction. This fact indicates that facial geometry and muscle movements are both beneficial for capturing depressive cues from facial activities. Therefore, in the following experiments, the inputs of models involved are $\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$ and $\mathbf{R}^{\mathrm{AU}}$.



(a) 331_1_05_None_gMLP (b) 240_2_16_Mild_gMLP (c) 329_1_24_Moderate_gMLP (d) 241_2_44_Severe_gMLP

(e) 331_1_05_None_DG (f) 240_2_16_Mild_DG (g) 329_1_24_Moderate_DG (h) 241_2_44_Severe_DG

Fig. 6. Visualization of results using two models with gMLP and DG modules for $\mathbf{R}^{\mathrm{AU}}$ of individuals with different depression levels. "320_3_05_None_gMLP" and "320_3_05_None_DG" refer to inputting $\mathbf{R}^{\mathrm{AU}}$ of 320_3 subject with the BDI-II score of 5 (None depression) into the gMLP+Linear and DG+Linear models for depression level estimation, respectively. The red box displays the discriminative parts.

TABLE IV

EXPERIMENTAL PERFORMANCE USING DIFFERENT MODELS AND INPUTS ON AVEC 2013 AND AVEC 2014 DEVELOPMENT SETS. "GPU" REFERS TO FKRS'S GUIDANCE ON AURS AS SHOWN BY THE RED ARROW IN FIG. 3 AND "GUP" REFERS TO AURS'S GUIDANCE ON FKRS AS SHOWN BY THE ORANGE ARROW IN FIG. 3. "CA" IS SHORT FOR CROSS-ATTENTION. "DG+MG+LINEAR" IS OUR CONSTRUCTED DEPRESSIONNET

Models	Inputs	AVEC 2013				AVEC 2014			
		RMSE (↓)	MAE (↓)	R^2 (↑)	SMAPE (↓)	RMSE (↓)	MAE (↓)	R^2 (↑)	SMAPE (↓)
DG+CA+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	7.76	6.28	0.58	66.25%	7.47	5.42	0.61	50.15%
DG+GPU+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	7.66	6.03	0.59	57.57%	7.15	5.13	0.64	49.81%
DG+GUP+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	6.99	5.22	0.66	51.34%	6.84	5.01	0.67	49.46%
DG+MG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	**6.21**	**4.88**	**0.73**	**50.28%**	**6.01**	**4.64**	**0.75**	**48.27%**

Furthermore, to illustrate the discriminative ability of DG module, we show the visualization results of $\mathbf{R}^{\mathrm{AU}}$ being processed by the models with gMLP and DG modules in Fig. 6. From this figure, it is easy to observe that our proposed DG module is more capable of perceiving differences in $\mathbf{R}^{\mathrm{AU}}$ of individuals with different depression levels compared to gMLP. This reason lies in the fact that the gMLP module performs residual addition on two sequences with different physical attributes, which reduces the model's discriminative ability. On the contrary, our DG module gates the input sequences from a spatiotemporal perspective to obtain effective global information and utilizes attention mechanisms to highlight content related to depression, while also ensuring the feasibility of operations between sequences.

2) The Role of MG Module: In order to take the advantages of FKRS and AURS, we propose the MG module to achieve interaction between different sequences. Therefore, in this section, we conduct various experiments to elucidate the effectiveness and necessity of MG module in improving the prediction accuracy of depression levels. The corresponding experimental results are given in Table IV.

From the experimental results in Table IV, we can observe that "DG+CA+Linear" has the lowest prediction accuracy. This reason is that CA is unable to maintain temporal attributes of the sequence during the process of aggregation, resulting in the loss of facial dynamic differences in individuals with different depression levels. Besides, all three guidance processes can bring improvements in prediction accuracy. Furthermore, one can find that the guidance of $\mathbf{R}^{\mathrm{AU}}$ on ($\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$) is more conducive to capturing differences in facial dynamics among individuals with different depression levels. The explanation for this result is that facial muscles are more able to reflect the impact of depression on subjects than FKs [44]. While, it is not difficult to find that our MG module enhances the ability to characterize depression cues by integrating facial muscle changes ($\mathbf{R}^{\mathrm{AU}}$) and spatial changes of facial keypoints ($\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$).

To further illustrate the effectiveness of MG module, we present the visualization results of different guidance processes in Fig. 7. From them, it is easy to see that the distinguish ability of the representation sequence has been further improved after the guidance process. Meanwhile, from Fig. 7 (e)-(h), it can be observed that the facial dynamic differences of individuals with different depression levels are more reflected in the FKs between the 40th and 50th points. In other words, dynamic changes in the corners of the eyes and mouth are more beneficial for predicting individual depression levels. Besides, from Fig. 7 (m)-(p), one can discover that AUs related to eyes and mouth are more helpful in predicting individual depression levels. This result is consistent with the conclusions of physiological studies [45], [46].

3) Significance Analysis of DG and MG Modules: From the results in Tables III and IV, it can be seen that our proposed DG and MG modules do indeed improve the prediction accuracy of depression levels. In this subsection, we use the Friedman test to examine whether the accuracy gain brought by those two modules are significant. The corresponding results are shown in Table V.

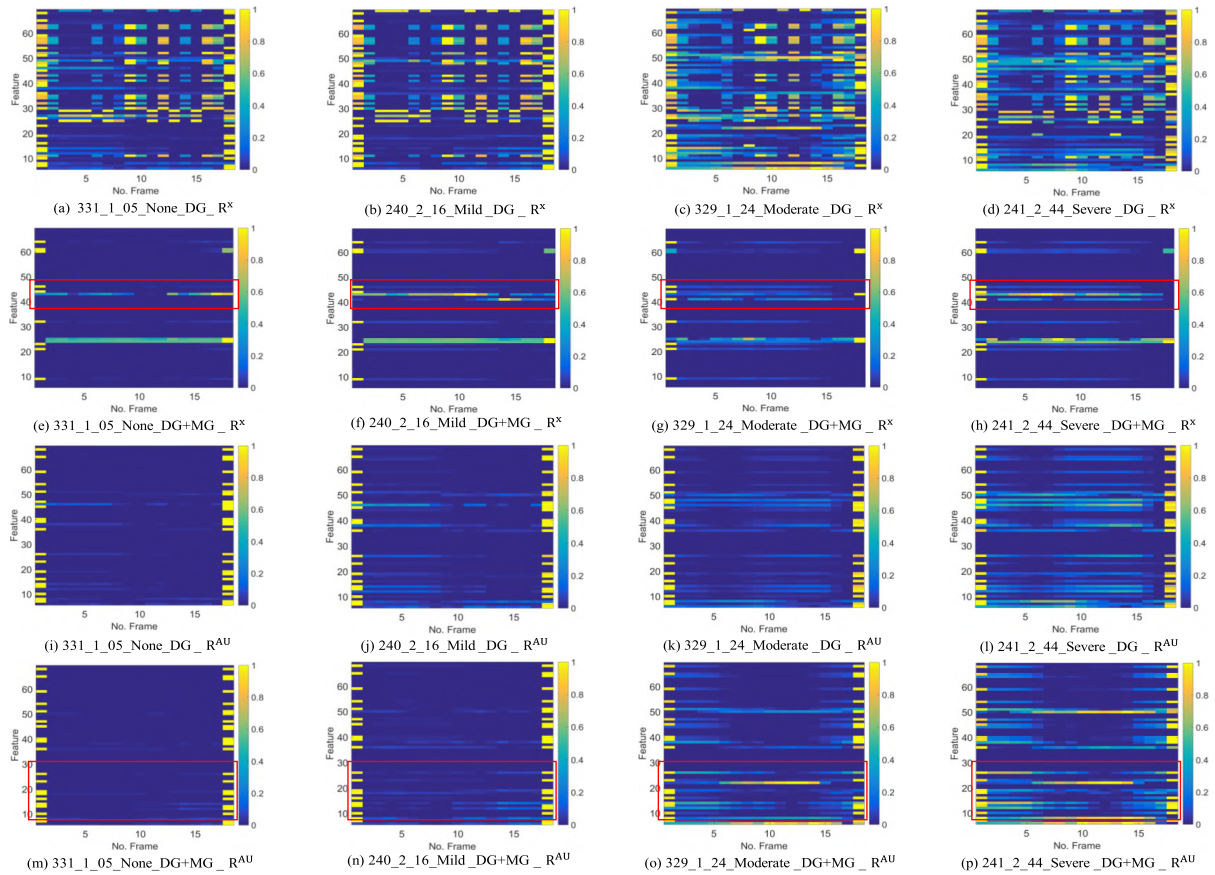


Fig. 7. Visualization of results using different guidance for $\mathbf{R}^{\mathrm{X}}$ and $\mathbf{R}^{\mathrm{AU}}$. "331_1_05_None_DG_$\mathbf{R}^{\mathrm{X}}$" and "320_3_05_None_DG+MG_$\mathbf{R}^{\mathrm{X}}$" refers to inputting $\mathbf{R}^{\mathrm{X}}$ of 320_3 subject with the BDI-II score of 5 (None depression) into the DG+Linear model and the DG+MG+Linear model for extracting processing result of $\mathbf{R}^{\mathrm{X}}$ and the guidance result of $\mathbf{R}^{\mathrm{AU}}$ on $\mathbf{R}^{\mathrm{X}}$, respectively. The red box displays the discriminative parts.

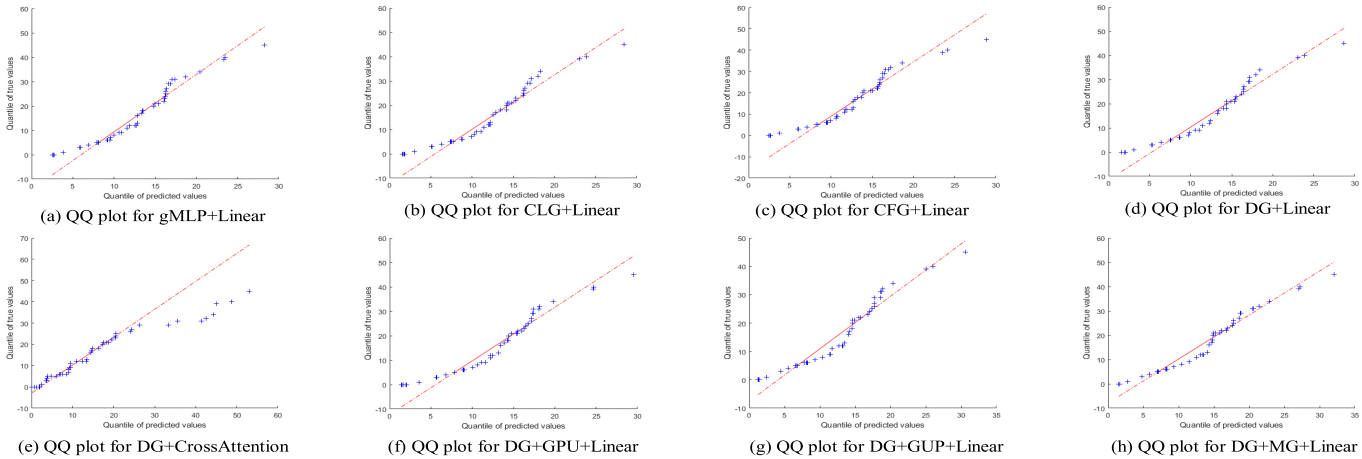
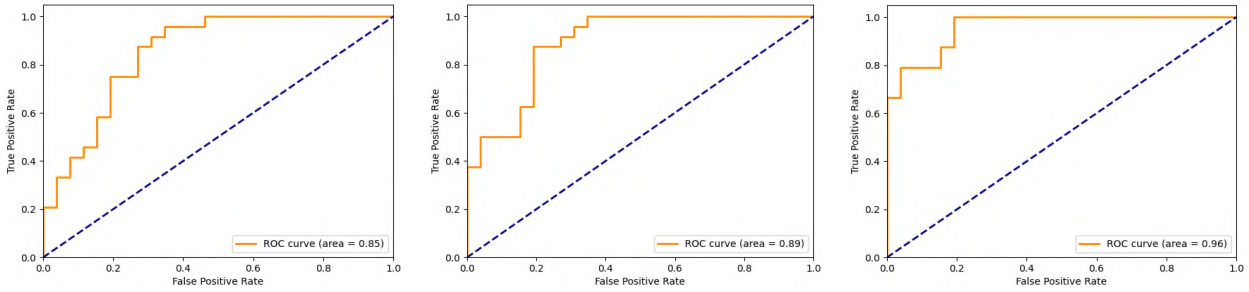


Fig. 8. The QQ plots for different models on the development set of AVEC 2013. The inputs of all models are $\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$ and $\mathbf{R}^{\mathrm{AU}}$. "DG+MG+Linear" is our constructed DepressionNet.

As shown in Table V, if the test parameter α in the Friedman test is set to 0.1, there is a significant difference in those eight models w.r.t the four metrics at a 90% confidence level. In other words, in a statistical sense, our proposed DG and MG modules are significant in improving the prediction accuracy of BDI-II scores. It is worth pointing that the p-values of RMSE, MAE, and R^2 are the same, as the ranking of prediction accuracy on AVEC 2013 and AVEC 2014 is consistent among different algorithms. Moreover, we have drawn the QQ plots of different models on the development set of AVEC 2013 in Fig. 8 to visually demonstrate the regression effects of different models. By comparing Fig. 8(a) and Fig. 8(d), as well as Fig. 8(e) and Fig. 8(h), it is not difficult to find that our proposed DG and MG modules can make the predicted BDI-II scores closer to the groundtruth.



(a) ROC curve for gMLP+Linear (b) ROC curve for DG+Linear (c) ROC curve for DG+MG+Linear

Fig. 9. ROC curves of two categories for different models on the development set of AVEC 2013. "DG+MG+Linear" is our constructed DepressionNet.

TABLE V

THE p-VALUE OF THE FRIEDMAN TEST FOR 8 MODELS WITH RESPECT TO 4 METRICS. "CA" IS SHORT FOR CROSS-ATTENTION. "DG+MG+LINEAR" IS OUR CONSTRUCTED DEPRESSIONNET

Models to be tested	Inputs	Metrics	p-values
gMLP+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	RMSE	0.0542
CLG+Linear			
CFG+Linear		MAE	0.0542
DG+Linear			
DG+CA+Linear		R^2	0.0542
DG+GPU+Linear			
DG+GUP+Linear		SMAPE	0.0952
DG+MG+Linear			

TABLE VI

PARAMETER SIZE AND COMPUTATIONAL COMPLEXITY OF DIFFERENT MODELS. "CA" IS SHORT FOR CROSS-ATTENTION. "DG+MG+LINEAR" IS OUR CONSTRUCTED DEPRESSIONNET

Models	Inputs	Prameter Size	FLOPs (M)
gMLP+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	11838	0.819
CLG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	22038	0.822
CFG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	15588	0.357
DG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	37374	1.187
DG+CA+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	40786	1.799
DG+GPU+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	40044	1.202
DG+GUP+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	39024	1.207
DG+MG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	41694	1.221

4) Model Complexity Analysis: In this subsection, we examine the parameter size and computation complexity of different models. The corresponding results are given in Table VI. From a parameter perspective, we can observe that our model requires the most parameters. The main reason for this result is that the DG module needs to extract cross-location and cross-frame gating results, while the MG module needs to generate the guidance masks for two types of sequences. Other models either do not need to obtain mixing gating results or only obtain the mixing result of a certain dimension. And those models do not need to generate guidance masks for different sequences. It is necessary to note that although our model has the largest number of parameters, we have ∼32 thousand short-term training samples obtained by segmenting long-term videos. Therefore, our model parameter size matches the number of training samples.

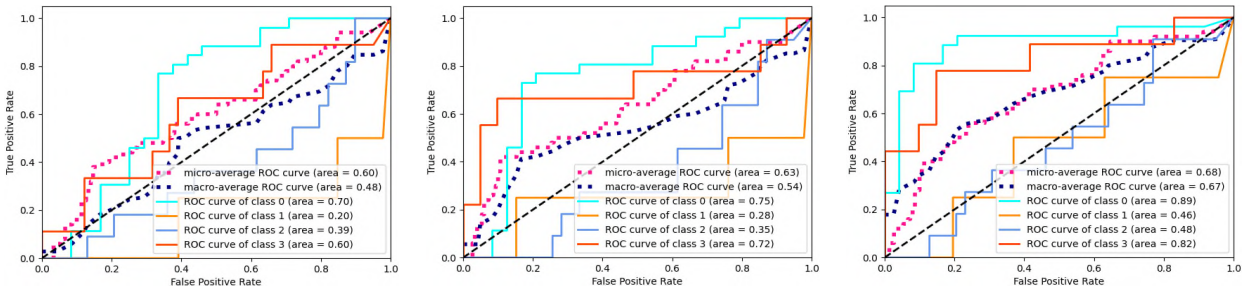
From a computational perspective, our model has the most FLOPs. This is because the DG module embeds cross-location and cross-frame gating results into the input sequence via attention mechanism to maintain the temporal attributes of the sequence. Besides, the MG module also implements guidance filtering operation between two types of sequences. Other models have less computational complexity as they do not require these operations. Furthermore, by comparing Tables V and VI, we can observe that although "DG+MG+Linear" involves the most parameter and computational complexity, it has achieved significant accuracy improvement in a statistical sense. This fact suggests that "DG+MG+Linear" does not suffer from overfitting and is effective for improving experimental performance in depression level prediction tasks.

5) Identification of Depression Severity Categories: As shown in Table I, the BDI-II score can classify subject's depression level into several categories. Therefore, in this subsection, we examine the recognition effects of different models on depression level categories. In our experiments, we conduct two categories and four categories experiments. In the experiment of two categories, "Mild", "Moderate" and "Severe" are merged into the one group termed as class 1, while "None" is treated as the other group termed as class 0. In the experiment of four categories, "None" (class 0), "Mild" (class 1), "Moderate" (class 2) and "Severe" (class 3) are each in one group. Table VII present the experimental results.

As shown in Table VII, "DG+MG+Linear" (namely DepressionNet) gains good performance in both two and four classification scenarios. Besides, one can observe that the accuracy of category four is lower than that of category two. This is because, in the AVEC 2013 and AVEC 2014 datasets, the number of samples for "Two categories" is comparable, while the number of samples for "Four categories" is very uneven. Furthermore, we also present the ROC curves for two categories and four categories in Fig. 9 and Fig. 10, respectively. From those figures, we can see that MG module has a better effect than DG module in improving the classification accuracy of depression levels. This fact indicates that there is indeed a complementary relationship between FKRS and AURS.

D. Comparisons With the Previous Works

In this section, we compare our method with current state-of-the-art works on the AVEC 2013 and AVEC 2014 test sets. The comparison of metric results for different methods is shown in Tables VIII and IX.



(a) ROC curve for gMLP+Linear (b) ROC curve for DG+Linear (c) ROC curve for DG+MG+Linear

Fig. 10. ROC curves of four categories for different models on the development set of AVEC 2013. "DG+MG+Linear" is our constructed DepressionNet.

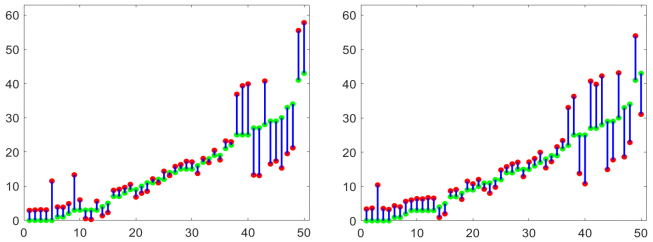
TABLE VII

CLASSIFICATION ACCURACY OF DEPRESSION LEVELS USING DIFFERENT MODELS ON THE DEVELOPMENT SETS OF AVEC 2013 AND AVEC 2014. "CA" IS SHORT FOR CROSS-ATTENTION. "DG+MG+LINEAR" IS OUR CONSTRUCTED DEPRESSIONNET

Number of categories	Models	Inputs	AVEC 2013			AVEC 2014		
			Recall (↑)	Precision (↑)	F1 (↑)	Recall (↑)	Precision (↑)	F1 (↑)
Two categories	gMLP+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.700	0.706	0.695	0.740	0.742	0.738
	CLG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.720	0.724	0.712	0.720	0.724	0.717
	CFG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.680	0.688	0.674	0.680	0.688	0.674
	DG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.740	0.742	0.738	0.720	0.724	0.717
	DG+CA+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.620	0.619	0.620	0.500	0.499	0.499
	DG+GPU+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.740	0.742	0.738	0.800	0.801	0.819
	DG+GUP+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.840	0.840	0.840	0.820	**0.821**	0.819
	DG+MG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	**0.860**	**0.861**	**0.860**	**0.820**	0.820	**0.819**
Four categories	gMLP+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.420	0.352	0.383	0.440	0.557	0.433
	CLG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.460	0.373	0.405	0.460	0.373	0.405
	CFG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.420	0.341	0.377	0.420	0.341	0.377
	DG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.460	0.385	0.411	0.480	0.553	0.441
	DG+CA+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.460	0.462	0.461	0.420	0.427	0.423
	DG+GPU+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.480	0.566	0.448	0.520	0.601	0.481
	DG+GUP+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	0.520	0.632	0.496	0.520	0.616	0.489
	DG+MG+Linear	$\mathbf{R}^{\mathrm{X}}$, $\mathbf{R}^{\mathrm{Y}}$, $\mathbf{R}^{\mathrm{AU}}$	**0.521**	**0.651**	**0.506**	**0.520**	**0.618**	**0.491**

In general, it is difficult to match the experimental performance of methods [4], [5], [47], [48], [49], [50] using hand-crafted features with those using neural networks [10], [12], [13], [35], [51], [52]. This is mainly because those hand-crafted features rely on the experience of designers and do not easily capture differences in facial dynamics among individuals with different depression levels. Moreover, the better prediction accuracy can be obtained by examining facial videos [20], [35], [51], [52] compared to facial images [7], [10]. This fact shows that depression cues are more reflected in facial dynamics. It can be observed that our model are not beyond those methods of [12], [13], and [20]. This is because the FKs and AUs used in our method cannot capture the facial texture changes of individuals with different depression levels, but our method can better protect personal privacy than them [12], [13], [20]. While, our method is superior to the methods [20], [21], [50] using FKs and AUs.

Furthermore, in Fig. 11, we present a comparison between the true and predicted values obtained using our Depression-MLP on the AVEC 2013 and AVEC 2014 test sets. From Fig. 11, we can see that our method performs well in predicting smaller BDI-II scores, but performs poorly in predicting larger BDI-II scores. The reason for this result is that the number of subjects with smaller BDI-II scores is greater than the number of subjects with larger BDI-II scores, which prevents the model from adequately capturing the facial dynamics characteristic of subjects with severe depression levels.



(a) Comparison of predicted and true values on the AVEC 2013 test set. (b) Comparison of predicted and true values on the AVEC 2014 test set.

Fig. 11. Comparison of predicted and true values on the test sets of AVEC 2013 (a) and AVEC 2014 (b). The horizontal and vertical axes represent the subject ID and BDI-II scores, respectively. The red solid dots and green solid dots represent the predicted and true scores, respectively. The blue vertical line represents the corresponding difference.

Besides, we calculate the mean RMSE (8.73 for AVEC 2013, 8.47 for AVEC 2014) and mean MAE (6.83 for AVEC 2013, 6.72 for AVEC 2014) for all methods in Tables VIII and IX, and find that the prediction results on the AVEC 2014 test set are better than those on the AVEC 2013 test set. This is because the AVEC 2014 dataset only contains videos of subjects completing a single task ("Northwind" or "FreeForm"), whereas the AVEC 2013 dataset places the videos collected when subjects complete all of the tasks together. Furthermore, under the same task, environmental

TABLE VIII

COMPARISON OF EXPERIMENTAL PERFORMANCE OF DEPRESSION LEVEL PREDICTION TASKS USING DIFFERENT METHODS ON THE AVEC 2013 TEST SET

Methods	Year	RMSE	MAE
Valstar et al. [47]	2013	13.61	10.88
Cummins et al. [48]	2013	10.45	–
Meng et al. [49]	2013	11.19	9.14
Wen et al. [53]	2015	10.27	8.22
Zhu et al. [54]	2017	9.82	7.58
Zhou et al. [7]	2017	8.28	6.20
He et al. [5]	2019	9.20	7.55
Niu et al. [4]	2019	9.17	6.97
Uddin et al. [40]	2020	8.93	7.04
Melo et al. [55]	2020	7.97	5.96
Jazaery et al. [39]	2021	9.28	7.37
Niu et al. [56]	2021	8.02	6.19
He et al. [34]	2021	8.39	6.59
Xu et al. [57]	2021	7.57	5.95
He et al. [58]	2022	8.46	6.83
Song et al. [21]	2022	8.10	6.16
Niu et al. [12]	2022	7.42	6.09
Shang et al. [10]	2023	8.20	6.38
Niu et al. [11]	2023	8.13	6.28
Casado et al. [59]	2023	8.01	6.43
Pan et al. [20]	2023	**7.26**	5.97
Melo et al. [60]	2023	7.55	6.24
Uddin et al. [13]	2023	7.32	5.90
Pan et al. [52]	2024	7.98	6.15
Melo et al. [35]	2024	7.66	6.14
Ours	–	**7.49**	**5.43**

TABLE IX

COMPARISON OF EXPERIMENTAL PERFORMANCE OF DEPRESSION LEVEL ESTIMATION TASKS USING DIFFERENT METHODS ON THE AVEC 2014 TEST SET

Methods	Year	RMSE	MAE
Valstar et al. [61]	2014	10.86	8.86
Espinosa et al. [50]	2014	9.84	8.46
Kaya et al. [62]	2014	10.26	8.20
Dhall et al. [63]	2015	8.91	7.08
Zhu et al. [54]	2017	9.55	7.47
He et al. [5]	2019	9.01	7.21
Zhou et al. [7]	2018	9.55	7.47
Niu et al. [4]	2019	9.10	7.19
Melo et al. [55]	2020	7.94	6.20
Jazaery et al. [39]	2021	9.20	7.22
Niu et al. [56]	2021	7.98	6.14
He et al. [34]	2021	8.30	6.51
Xu et al. [57]	2021	7.18	5.86
Song et al. [21]	2022	7.15	5.95
Uddin et al. [40]	2022	8.78	6.86
Niu et al. [12]	2022	7.39	5.87
He et al. [58]	2022	8.42	6.78
Shang et al. [10]	2023	7.84	6.08
Niu et al. [11]	2023	8.07	6.14
Casado et al. [59]	2023	8.49	6.57
Pan et al. [20]	2023	7.30	5.99
Melo et al. [60]	2023	7.65	6.06
Uddin et al. [13]	2023	**6.98**	5.75
Pan et al. [52]	2024	7.75	6.00
Melo et al. [35]	2024	7.50	5.69
Ours	–	**7.27**	**5.63**

variables are more easily controlled and the models are more likely to capture differences in facial dynamics among individuals with different depression levels.

To demonstrate the robustness of our method to different depression scale scores, we also conduct experiments on the DAIC-WOZ dataset. The corresponding results are shown in Table X. By comparing with other current works, we can find that our DepressionMLP model still achieves good prediction accuracy. This fact suggests that the performance of our method is stable in capturing facial dynamic differences among individuals with different depression levels, even when subjects are labeled using different depression scale scores.

TABLE X

COMPARISON OF EXPERIMENTAL PERFORMANCE OF DEPRESSION LEVEL ESTIMATION TASKS USING DIFFERENT METHODS ON THE DAIC-WOZ TEST SET

Methods	Years	RMSE	MAE
Valstar et al. [64]	2016	6.97	6.12
Williamson et al. [65]	2016	6.54	5.33
Nasir et al. [66]	2016	7.86	6.48
Song et al. [24]	2018	5.84	4.37
Du et al. [18]	2019	5.78	4.64
Rathi et al. [67]	2019	5.98	4.64
Qureshi et al. [68]	2019	6.53	5.05
Zhang et al. [69]	2022	6.45	4.97
Rasipuram et al. [70]	2022	5.76	4.83
Wei et al. [71]	2022	8.06	6.17
Chen et al. [72]	2022	5.11	4.38
Rumahorbo et al. [73]	2023	6.27	5.39
Fang et al. [74]	2023	5.44	4.12
Shu et al. [75]	2023	5.14	**3.97**
Ours	–	**5.03**	**4.11**

V. CONCLUSION

Physiological studies have shown that depression can cause patients to exhibit facial movements that are different from those of healthy individuals. Therefore, we construct a DepressionMLP architecture to predict the depression level using FKRS and AURS for protecting individual privacy. In our model, the DG module is proposed to embed cross-frame and cross-location gating results into the input sequence through the attention mechanism, thus maintaining the spatiotemporal attributes of sequences and ensuring the additivity of residual connections in a physical sense. Moreover, our proposed MG module can achieve the interaction among various sequences through mutual guidance processes and enhance the model's discriminative ability. The experimental results on the AVEC 2013, AVEC 2014 and DAIC-WOZ depression databases demonstrate the effectiveness of our approach. In the future, we will construct a dataset with clear interaction task boundaries and examine the movement patterns of FKs and AUs for depressed and healthy individuals under different interaction tasks.

REFERENCES

[1] C. Otte et al., "Major depressive disorder," *Nature Rev. Disease Primers*, vol. 2, no. 1, pp. 1–20, 2016.

[2] A. D. Sirota and G. E. Schwartz, "Facial muscle patterning and lateralization during elation and depression imagery," *J. Abnormal Psychol.*, vol. 91, no. 1, pp. 25–34, 1982.

[3] E. Finzi and N. E. Rosenthal, "Emotional proprioception: Treatment of depression with afferent facial feedback," *J. Psychiatric Res.*, vol. 80, pp. 93–96, Sep. 2016.

[4] M. Niu, J. Tao, and B. Liu, "Local second-order gradient cross pattern for automatic depression detection," in *Proc. 8th Int. Conf. Affect. Comput. Intell. Interact. Workshops Demos (ACIIW)*, Sep. 2019, pp. 128–132.

[5] L. He, D. Jiang, and H. Sahli, "Automatic depression analysis using dynamic facial appearance descriptor and Dirichlet process Fisher encoding," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1476–1486, Jun. 2019.

[6] Y. Kang et al., "Deep transformation learning for depression diagnosis from facial images," in *Proc. Chin. Conf. Biometric Recognit.*, vol. 10568, Shenzhen, China. Springer, 2017, p. 13.

[7] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Trans. Affect. Comput.*, vol. 11, no. 3, pp. 542–552, Jul. 2020.

[8] X. Zhou, Z. Wei, M. Xu, S. Qu, and G. Guo, "Facial depression recognition by deep joint label distribution and metric learning," *IEEE Trans. Affect. Comput.*, vol. 13, no. 3, pp. 1605–1618, Jul. 2022.

[9] X. Li, W. Guo, and H. Yang, "Depression severity prediction from facial expression based on the DRR_DepressionNet network," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2020, pp. 2757–2764.

[10] Y. Shang et al., "LQGDNet: A local quaternion and global deep network for facial depression recognition," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2557–2563, Dec. 2023.

[11] M. Niu, Z. Zhao, J. Tao, Y. Li, and B. W. Schuller, "Dual attention and element recalibration networks for automatic depression level prediction," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 1954–1965, May 2023.

[12] M. Y. Niu, Z. P. Zhao, J. H. Tao, Y. Li, and B. W. Schuller, "Selective element and two orders vectorization networks for automatic depression severity diagnosis via facial changes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 8065–8077, Nov. 2022.

[13] M. A. Uddin, J. B. Joolee, and K.-A. Sohn, "Deep multi-modal network based automated depression severity estimation," *IEEE Trans. Affect. Comput.*, vol. 14, no. 3, pp. 2153–2167, Jun. 2023.

[14] M. Niu, L. He, Y. Li, and B. Liu, "Depressioner: Facial dynamic representation for automatic depression level prediction," *Expert Syst. Appl.*, vol. 204, Oct. 2022, Art. no. 117512.

[15] L. He, P. Tiwari, C. Lv, W. Wu, and L. Guo, "Reducing noisy annotations for depression estimation from facial images," *Neural Netw.*, vol. 153, pp. 120–129, Sep. 2022.

[16] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri, "Comparison of beck depression inventories-IA and-II in psychiatric outpatients," *J. Person. Assess.*, vol. 67, no. 3, pp. 588–597, 1996.

[17] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[18] Z. Du, W. Li, D. Huang, and Y. Wang, "Encoding visual behaviors with attentive temporal convolution for depression prediction," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2019, pp. 1–7.

[19] L. Yang, D. Jiang, and H. Sahli, "Integrating deep and shallow models for multi-modal depression analysis—Hybrid architectures," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 239–253, Jan. 2021.

[20] Y. Pan, Y. Shang, Z. Shao, T. Liu, G. Guo, and H. Ding, "Integrating deep facial priors into landmarks for privacy preserving multimodal depression recognition," *IEEE Trans. Affect. Comput.*, early access, Jul. 17, 2023, doi: 10.1109/TAFFC.2023.3296318.

[21] S. Y. Song, S. Jaiswal, L. L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 829–844, Apr./Jun. 2022.

[22] Z. S. Syed, K. Sidorov, and D. Marshall, "Depression severity prediction based on biomarkers of psychomotor retardation," in *Proc. 7th Annu. Workshop Audio/Visual Emotion Challenge*, Oct. 2017, pp. 37–43.

[23] A. Ray, S. Kumar, R. Reddy, P. Mukherjee, and R. Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Proc. 9th Int. Audio/Vis. Emotion Challenge Workshop*, Oct. 2019, pp. 81–88.

[24] S. Song, L. Shen, and M. Valstar, "Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 158–165.

[25] M. Muzammel, H. Salam, and A. Othmani, "End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis," *Comput. Methods Programs Biomed.*, vol. 211, Nov. 2021, Art. no. 106433.

[26] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to MLPs," in *Proc. Conf. Neural Inf. Proces. Syst. (NeurIPS)*, 2021, pp. 9204–9215.

[27] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[28] I. Tolstikhin et al., "MLP-mixer: An all-MLP architecture for vision," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.

[29] Z. Wang et al., "DynaMixer: A vision MLP architecture with dynamic mixing," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 22691–22701.

[30] C. Tang, Y. Zhao, G. Wang, C. Luo, W. Xie, and W. Zeng, "Sparse MLP for image recognition: Is self-attention really necessary?" in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 2344–2351.

[31] Y. Dong, Q. Liu, B. Du, and L. Zhang, "Weighted feature fusion of convolutional neural network and graph attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 1559–1572, 2022.

[32] J. Lu, H. Wan, P. Li, X. Zhao, N. Ma, and Y. Gao, "Exploring high-order spatio-temporal correlations from skeleton for person re-identification," *IEEE Trans. Image Process.*, vol. 32, pp. 949–963, 2023.

[33] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "CAVER: Cross-modal view-mixed transformer for bi-modal salient object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 892–904, 2023.

[34] L. He, J. C.-W. Chan, and Z. Wang, "Automatic depression recognition using CNN with attention mechanism from videos," *Neurocomputing*, vol. 422, pp. 165–175, Jan. 2021.

[35] W. C. de Melo, E. Granger, and M. B. Lopez, "Facial expression analysis using decomposed multiscale spatiotemporal networks," *Expert Syst. Appl.*, vol. 236, Feb. 2024, Art. no. 121276.

[36] M. Niu, J. Tao, B. Liu, J. Huang, and Z. Lian, "Multimodal spatiotemporal representation for automatic depression level detection," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 294–307, Jan. 2023.

[37] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[38] J. Gratch et al., "The distress analysis interview corpus of human and computer interviews," in *Proc. Int. Conf. Lang. Resour. Eval.*, 2014, pp. 3123–3128.

[39] M. Al Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Trans. Affect. Comput.*, vol. 12, no. 1, pp. 262–268, Jan. 2021.

[40] M. A. Uddin, J. B. Joolee, and Y. Lee, "Depression level prediction using deep spatiotemporal features and multilayer bi-LTSM," *IEEE Trans. Affect. Comput.*, vol. 13, no. 2, pp. 864–870, Apr. 2022.

[41] M. L. Cañellas, C. Á. Casado, L. Nguyen, and M. B. López, "Depression recognition from facial videos: Preprocessing and scheduling choices hide the architectural contributions," *Electron. Lett.*, vol. 59, no. 20, Oct. 2023, Art. no. e12992.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[43] R. Flores, M. L. Tlachac, A. Shrestha, and E. Rundensteiner, "Temporal facial features for depression screening," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2022, pp. 488–493.

[44] D. S. B. A. Hamid, S. B. Goyal, and P. Bedi, "Integration of deep learning for improved diagnosis of depression using EEG and facial features," *Mater. Today, Proc.*, vol. 80, pp. 1965–1969, Jan. 2023.

[45] J. L. Stewart, J. A. Coan, D. N. Towers, and J. J. B. Allen, "Frontal EEG asymmetry during emotional challenge differentiates individuals with and without lifetime major depressive disorder," *J. Affect. Disorders*, vol. 129, nos. 1–3, pp. 167–174, Mar. 2011.

[46] C. Bourke, K. Douglas, and R. Porter, "Processing of facial emotion expression in major depression: A review," *Austral. New Zealand J. Psychiatry*, vol. 44, no. 8, pp. 681–696, Aug. 2010.

[47] M. Valstar et al., "AVEC 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2013, pp. 3–10.

[48] N. Cummins, J. Joshi, A. Dhall, V. Sethu, R. Goecke, and J. Epps, "Diagnosis of depression by behavioural signals: A multimodal approach," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2013, pp. 11–20.

[49] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proc. 3rd ACM Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2013, pp. 21–30.

[50] H. P. Espinosa, H. J. Escalante, L. Villaseñor-Pineda, M. Montes-y-Gómez, D. Pinto-Avedaño, and V. Reyez-Meza, "Fusing affective dimensions and audio-visual features from segmented video for depression recognition: INAOE-BUAP's participation at AVEC'14 challenge," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, Nov. 2014, pp. 49–55.

[51] S. Zhang et al., "MTDAN: A lightweight multi-scale temporal difference attention networks for automated video depression detection," *IEEE Trans. Affect. Comput.*, early access, Sep. 25, 2023, doi: 10.1109/TAFFC.2023.3312263.

[52] Y. Pan et al., "Spatial–temporal attention network for depression recognition from facial videos," *Expert Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121410.

[53] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 7, pp. 1432–1441, Jul. 2015.

[54] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 578–584, Oct. 2018.

[55] W. C. de Melo, E. Granger, and M. B. Lopez, "Encoding temporal information for automatic depression recognition from facial analysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1080–1084.

[56] M. Niu, J. Tao, and B. Liu, "Multi-scale and multi-region facial discriminative representation for automatic depression level prediction," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1325–1329.

[57] J. Xu, S. Song, K. Kusumam, H. Gunes, and M. Valstar, "Two-stage temporal modelling framework for video-based depression recognition using graph representation," 2021, *arXiv:2111.15266*.

[58] L. He, C. Guo, P. Tiwari, H. M. Pandey, and W. Dang, "Intelligent system for depression scale estimation with facial expressions and case study in industrial intelligence," *Int. J. Intell. Syst.*, vol. 37, no. 12, pp. 10140–10156, Dec. 2022.

[59] C. Á. Casado, M. L. Cañellas, and M. B. López, "Depression recognition using remote photoplethysmography from facial videos," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3305–3316, Jan. 2023.

[60] W. C. de Melo, E. Granger, and M. B. López, "MDN: A deep maximization-differentiation network for spatio-temporal depression detection," *IEEE Trans. Affect. Comput.*, vol. 14, no. 1, pp. 578–590, Jan. 2023.

[61] M. Valstar et al., "AVEC 2014: 3D dimensional affect and depression recognition challenge," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, Nov. 2014, pp. 3–10.

[62] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble CCA for continuous emotion prediction," in *Proc. 4th Int. Workshop Audio/Visual Emotion Challenge*, Nov. 2014, pp. 19–26.

[63] A. Dhall and R. Goecke, "A temporally piece-wise Fisher vector approach for depression analysis," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 255–259.

[64] M. Valstar et al., "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2016, pp. 3–10.

[65] J. R. Williamson et al., "Detecting depression using vocal, facial and semantic communication cues," in *Proc. 6th Int. Workshop Audio/Visual Emotion Challenge*, Oct. 2016, pp. 11–18.

[66] M. Nasir, A. Jati, P. G. Shivakumar, S. N. Chakravarthula, and P. Georgiou, "Multimodal and multiresolution depression detection from speech and facial landmark features," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge*, Oct. 2016, pp. 43–50.

[67] S. Rathi, B. Kaur, and R. K. Agrawal, "Enhanced depression detection from facial cues using univariate feature selection techniques," in *Proc. Int. Conf. Pattern Recognit. Mach. Intell.* Cham, Switzerland: Springer, 2019, pp. 22–29.

[68] S. A. Qureshi, M. Hasanuzzaman, S. Saha, and G. Dias, "The verbal and non verbal signals of depression—Combining acoustics, text and visuals for estimating depression level," 2019, *arXiv:1904.07656*.

[69] W. Zhang, "Biomedical engineering application: Disease diagnosis and treatment," Ph.D. thesis, Univ. Alberta, Edmonton, AB, Canada, 2022.

[70] S. Rasipuram, J. H. Bhat, A. Maitra, B. Shaw, and S. Saha, "Multimodal depression detection using task-oriented transformer-based embedding," in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2022, pp. 01–04.

[71] P.-C. Wei, K. Peng, A. Roitberg, K. Yang, J. Zhang, and R. Stiefelhagen, "Multi-modal depression estimation based on sub-attentional fusion," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 623–639.

[72] M. Chen, X. Xiao, B. Zhang, X. Liu, and R. Lu, "Neural architecture searching for facial attributes-based depression recognition," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 877–884.

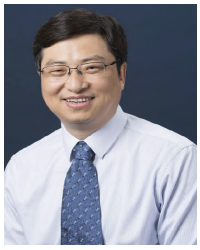
[73] B. N. Rumahorbo, B. Pardamean, and G. N. Elwirehardja, "Exploring recurrent neural network models for depression detection through facial expressions: A systematic literature review," in *Proc. 6th Int. Conf. Comput. Informat. Eng. (IC2IE)*, Sep. 2023, pp. 209–214.

[74] M. Fang, S. Peng, Y. Liang, C.-C. Hung, and S. Liu, "A multimodal fusion model with multi-level attention mechanism for depression detection," *Biomed. Signal Process. Control*, vol. 82, Apr. 2023, Art. no. 104561.

[75] T. Shu, F. Zhang, and X. Sun, "Gaze behavior based depression severity estimation," in *Proc. IEEE 4th Int. Conf. Pattern Recognit. Mach. Learn. (PRML)*, Aug. 2023, pp. 313–319.

Mingyue Niu received the Ph.D. degree from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA). At present, he is currently a Lecturer with the School of Information Science and Engineering, Yanshan University. He had published the papers in IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, ICASSP, and INTERSPEECH. His research interests include affective computing and multimodal learning.

Ya Li (Member, IEEE) is currently an Associate Professor with Beijing University of Posts and Telecommunications (BUPT). She has published more than 70 papers in journals and conferences, including *Speech Communication*, INTERSPEECH, ICASSP, and *ACII*. Her research interests include speech synthesis, affective computing, and multimodal interaction. She won the first prize in Science and Technology Progress Award of China Electronics Society in 2018 and the second prize in Beijing Science and Technology Award in 2014.



Jianhua Tao (Senior Member, IEEE) is currently a Professor with the Department of Automation, Tsinghua University. He has directed many national projects, including "863" and the National Natural Science Foundation of China. He has published more than eighty papers in journals and proceedings, including IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, ICASSP, and INTERSPEECH. His research interests include speech synthesis, affective computing, and pattern recognition. He serves as a Steering Committee Member for IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the Chair or a Program Committee Member for major conferences, including ICPR and INTERSPEECH. He is a Winner of the National Science Fund for Distinguished Young Scholars.

Xiuzhuang Zhou (Member, IEEE) is currently a Professor with the School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China. He has authored more than 40 scientific papers in peer-reviewed journals and conferences, including some top venues, such as IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, CVPR, and ACM Multimedia. His research interests include computer vision, pattern recognition, and machine learning. He is serving as an Associate Editor for *Neurocomputing*.

Björn W. Schuller (Fellow, IEEE) received the Ph.D. degree from Technische Universität München, Germany. He is currently a Professor of artificial intelligence with the Department of Computing, Imperial College London, U.K., and a Full Professor and the Head of the Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany. He (co)authored five books and more than 1 000 publications in peer-reviewed books, journals, and conference proceedings. He is a fellow of ISCA and a Senior Member of ACM. He is the Field Chief Editor of *Frontiers in Digital Health*, the President-Emeritus of AAAC, and the Golden Core Awardee of the IEEE Computer Society.